Fuzzy Systems and Soft Computing
ISSN : 1819-4362

MULTI-OBJECTIVE WRAPPER BASED FEATURE SELECTION USING GENETIC ALGORITHM FOR MEDICAL DATASETS

Anitha G Assistant Professor, Department of Data Analytics, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India. ganithamca@gmail.com
Rosiline Jeetha B Associate Professor and Head, Department of Computer Science, Nirmala College for Women, Coimbatore, Tamil Nadu, India.jeethasekar@gmail.com

Abstract:

In the era of big data, an enormous amount of high-dimensional data is being produced. High-dimensional data poses a significant challenge to current learning techniques, i.e., the curse of dimensionality. Due to high dimensionality, learning models tend to over fit, which degenerates the performance. Feature selection is crucial for enhancing prediction models' performance since it identifies important features and eliminates those that are redundant. The primary goal of feature selection is to identify the optimal subset of features, by reducing dimensionality and improving accuracy.The feature selection (FS) process is a multi-objective optimization problem that aims to maximize classification performance while minimizing the number of features. A genetic algorithm is a one of the best optimization approach for feature selection it can effectively search through a large feature space. This study focuses on identifying the most informative subset of features to improve the prediction accuracy in healthcare datasets. It emphasizes the significance of feature selection in optimizing the performance of predictive models. A multi-objective wrapper-based feature selection technique using a Genetic Algorithm (GA) is implemented with SVM, KNN, XGBoost, and Random Forest classifiers in this study. The findings show that the XGBoost outperforms other classifier algorithms.

Keywords:

Genetic Algorithm, Feature Selection, Meta heuristic, wrapper approach.

I. Introduction:

Large amounts of data are present in real-world applications, data handling is a challenging and difficult task. Datasets typically consist of numerous attributes or features. However, not all features are essential for extracting meaningful information from datasets. Some may be irrelevant or redundant, potentially degrading algorithm performance. Feature selection reduces data dimensionality by eliminating irrelevant and redundant features. This enhances learning speed, simplifies the model, and improves algorithm performance[1]. Feature selection has been an active research area in data mining for decades and is extensively used in fields such as medical diagnosis, document classification, genomic analysis, object recognition, and complex technological process modeling. Feature selection can be applied to ultrahigh-dimensional data, streaming data, multi-task data, and multi-source data.

Feature selection is among the most challenging tasks in machine learning and data mining. Given a set of n features, there exist 2^n possible subsets, from which the best one must be selected. As n increases, evaluating the model’s performance for each subset becomes computationally infeasible[2]. Therefore, several methodologies have been proposed to address this challenge.

The process of feature selection consists of three main steps: (1) Subset Discovery -Creating subsets of features from the original dataset using certain search techniques. (2) Subset Evaluation: Determining the significance of the features in the candidate set. Depending on the importance, specific characteristics from the candidate set may be added or removed from the chosen feature set. (3) Stopping Criteria: This criterion is used to assess whether the current selection of features is sufficient [2].

Based on the evaluation criteria, feature selection algorithms are generally categorized as: 1) Filter Approach 2) Wrapper Approach and 3) Embedded Approach. Filter Approach: To evaluate the utility of features, Filter methods assess the general characteristics of data without using any classification algorithm. Relief, Fisher score, Mutual Information, correlation and Information Gain methods are most well-known algorithms in the filter model[3]. Filter methods are typically fast and suitable for large-scale datasets. Wrapper Approach: In contrast, wrapper methods require a predefined classification algorithm to assess feature subsets and determine the optimal set. This methods are computationally expensive and slower than the filter Approach[4]. Embedded Approach: The integration of both filter and wrapper approaches into a single process is known as the embedded approach.

Evolutionary algorithms approaches for the feature selection process are considered in this study. This is a subset of evolutionary computation. It is a population-based metaheuristic algorithm. The objective is to propose a Genetic Algorithm with a wrapper-based feature selection approach, where different classifiers are used for evaluation. This paper is organized as follows. Section II discusses the review of literature, section III covers Meta heuristics Algorithm concepts. Section V discusses the experimental results and finally section V concludes the paper.

II. Review of Literature:

This section elaborates on the literature review work related to the various techniques to solve the feature selection problem. Asgarnezhad et.al (2021) created a multi-objective genetic algorithm for feature selection to classify the text. Evolutionary algorithms are suitable for providing optimized solutions to feature selection problems[5]. Xue et.al (2016) proposed a feature selection technique using evolutionary algorithms[2]. An adaptive PSO for feature selection approach was introduced by Li et.al (2016). Strategies to improve classification performance and reduce computation time were proposed[6]. Ant colony optimization approach using wrapper-filter model for feature selection is proposed by Ghosh et.al (2020)[7].

Neesha Jothi et al.,(2019) applied wrapper-based feature selection using the Genetic Algorithm (GA) and used SVM classifier for medical datasets.GA enhances the classification accuracy will help medical practitioners in making better diagnoses for patients[8].Ravi Kumar et al (2014)applied SVM and Binary Coded Genetic Algorithm for effective FS in medical data classification.The GA-SVM gives better results than the traditional SVM[9]. Karegowda et al. (2010) used a wrapper approach for feature subset selection using a genetic algorithm as the random search technique, They integrated four different classifiers as the subset evaluation mechanism, and evaluated the performance of the selected feature subsets on four standard datasets[10].

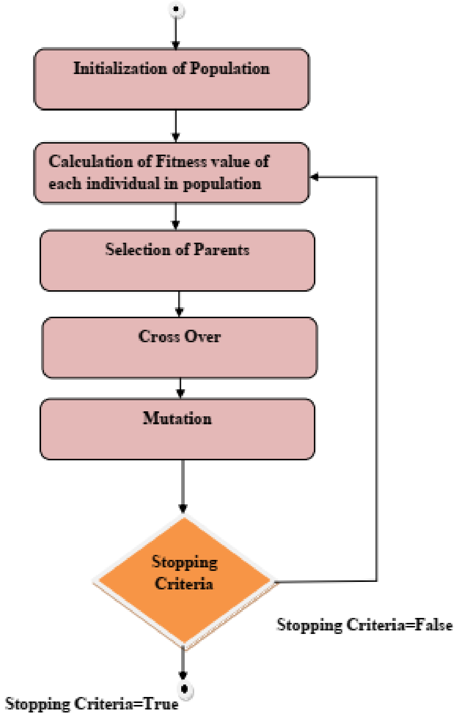
N. Hewahi and Alashqar (2015) proposed a novel approach that integrates Genetic Algorithms (GA) with a Correlation Ranking Filter (CRF) wrapper to select a small subset of features from a large set of geospatial features extracted from satellite imagery. This method aims to enhance the accuracy and efficiency of object recognition classifiers, including Neural Networks, K-Nearest Neighbor, and Decision Trees[11]. Jinjie Huang (2007) proposed a hybrid Genetic Algorithm for feature selection that incorporates mutual information in a two-stage optimization process to achieve both high global predictive accuracy and efficient local search[12]. Prokopis K. et al.(2017) described an approach for optimizing feature selection in decision tree classification using a Genetic Algorithm wrapper method, which helps reduce overfitting and improve accuracy, particularly for large datasets[13].

III. Meta Heuristic Algorithm:

A metaheuristic algorithm is an optimization technique that effectively explores and exploits search space to find solutions to hard problems. Metaheuristic algorithms can be applied to identify an optimal or near optimal subset of features in feature selection[14]. Algorithms such as genetic algorithms, particle swarm optimization, and simulated annealing perform evaluations of different feature combinations of a subsets based on a fitness or objective function to identify the best subset of features. Metaheuristic algorithms are classified into two types. single-solution-based approach, referred to as local search techniques, and population-based approach it belongs to global search techniques. The local search search algorithms start with a solution and try to improvise until the stop condition is satisfied. The current region of the problem's search space is extensively explored by this algorithm. This search method typically results in a local optimization issue. Population-based metaheuristic algorithms utilize stochastic techniques to explore the search space and optimize solutions globally. Population-based algorithms draw inspiration from biological processes involving living organisms and the social behaviors of animals in nature[15].These algorithms effectively explore the search space and typically produce feasible solutions.

Genetic Algorithm:

GA falls under a population-based metaheuristic algorithm. The Genetic Algorithm (GA) is based on Charles Darwin’s theory of evolution, replicates natural selection by selecting the fittest individuals for reproduction to create future generations[6].GA is a component of the metaheuristic algorithm, which optimizes the solutions to various computer science problems through feature selection. Figure 1 displays the GA flow graph.



A genetic algorithm mainly composed of following steps: Initialization/Population, reproduction, crossover, selection and mutation.

1. Initialization/Population: Initializing the population is the first step in genetic algorithms for feature selection. Every member of the population represents a possible feature subset, which is usually represented by a binary string with 1 denoting a feature that is chosen and 0 denoting one that is not. The initial population is created randomly, with each individual choosing a distinct subset of features. The population's size is a parameter that influences the algorithm's capability for exploration.

2. Fitness: Reproduction uses an objective function to randomly select chromosomes from a population and chooses good strings (a subset of input attributes). Individuals are chosen using the objective functions and the survival of the fittest principle [16].

3. Crossover: In a genetic algorithm, crossover is the most crucial stage. Crossover, or recombination, is used to combine the genes of two parents to produce offspring. Crossover aims to explore the new areas in the search space while inheriting the strengths of both parents.

4. Selection: Selection chooses which individuals of the population will reproduce and produce their next generation. The selection process is based on an individual's rank after they have been ranked according to their level of fitness.

5. Mutation: To preserve population diversity, mutations make small arbitrary modifications to an individual's chromosomes. A few chromosomal bits are randomly flipped by this operator [9].In an effort to produce a better string, mutation modifies a string locally. Mutation improves diversification and addresses the problem of premature convergence.

6. Termination: The GA stops execution once a predefined termination criterion is satisfied, such as reaching the maximum number of generations[17].

IV. Experimental setups, UCI benchmark datasets, and parameter tuning:

An Intel Core i7 7th generation processor running at 2.7 GHz, a 500GB hard drive, 16 GB of RAM, and Microsoft Windows 10 OS are used for the experiments. The Python 3.9 language and its additional libraries are used to evaluate the proposed model. This study experiments with 12 benchmark datasets from the UC Irvine Machine Learning Repository. The Dataset details are shown in Table 1. Parameter settings are shown in Table 2. The average classification accuracy of the different classifiers are shown in Table 3.

Table 1. Dataset Description

Dataset Name	**Attributes Count**	**Instance Count**
Breast cancer	30	569
PIMA Indians Diabetes Dataset	8	768
Heart Disease UCI dataset	14	303
Colon Dataset	2000	62
Prostate Cancer dataset	10509	102
CNS (Central Nervous System) Tumor dataset	7129	60
Ovarian dataset	15,154	253
Mammographic Mass Dataset	6	961
Dermatology dataset	34	366
Thoracic Surgery dataset	17	470

Table 2 .Parameter setup

Parameter Name	**Value**
Population Size	20
Number of Generations	40
Crossover Probability	0.5
Mutation Probability	0.2
Tournament Size	3

Table. 3 Average Classification Accuracy of Classifiers

KNN	0.8258
SVM	0.8445
XGBoost	0.8523
Random Forest	0.8363

V. CONCLUSION & FUTURE WORK:

This proposed method,used medical datasets, showed significant improvement in classification accuracy through feature selection.The XGBOOST Classifier (85.23%) produced the greatest test accuracy out of the 4 classifiers used to evaluate the model. The findings demonstrate that choosing the right characteristics can enhance classification accuracy. This study shows that a genetic algorithm performs well for feature selection and could be applied to other medical datasets.Even though the feature selection techniques had good prediction accuracy, GA required a long time to compute and there was evidence that the system was stuck in local optima. To overcome these challenges, our future work will focus on improving predictions through hybrid feature selection techniques.

References:

[1]. Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, “Approaches to Multi-Objective Feature Selection: A Systematic Literature Review,” *IEEE Access*, vol. 8, pp. 125076–125096, 2020, doi: 10.1109/ACCESS.2020.3007291.

[2]. B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A Survey on Evolutionary Computation Approaches to Feature Selection,” *IEEE Trans. Evol. Computat.*, vol. 20, no. 4, pp. 606–626, Aug. 2016, doi: 10.1109/TEVC.2015.2504420.

[3]. S. Kashef and H. Nezamabadi-pour, “An advanced ACO algorithm for feature subset selection,” *Neurocomputing*, vol. 147, pp. 271–279, Jan. 2015, doi: 10.1016/j.neucom.2014.06.067.

[4]. H. M.Harb and A. S. Desuky, “Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization,” *IJCA*, vol. 104, no. 5, pp. 14–17, Oct. 2014, doi: 10.5120/18197-9118.

[5]. R. Asgarnezhad, S. A. Monadjemi, and M. Soltanaghaei, “An application of MOGW optimization for feature selection in text classification,” *J Supercomput*, vol. 77, no. 6, pp. 5806–5839, Jun. 2021, doi: 10.1007/s11227-020-03490-w.

[6]. Y. Xue, B. Xue, and M. Zhang, “Self-Adaptive Particle Swarm Optimization for Large-Scale Feature Selection in Classification,” *ACM Trans. Knowl. Discov. Data*, vol. 13, no. 5, pp. 1–27, Oct. 2019, doi: 10.1145/3340848.

[7]. M. Ghosh, R. Guha, R. Sarkar, and A. Abraham, “A wrapper-filter feature selection technique based on ant colony optimization,” *Neural Comput & Applic*, vol. 32, no. 12, pp. 7839–7857, Jun. 2020, doi: 10.1007/s00521-019-04171-3.

[8]. N. Jothi, W. Husain, N. Abdul Rashid, and S. M. Syed-Mohamad, “Feature Selection Method using Genetic Algorithm for Medical Dataset,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 9, no. 6, pp. 1907–1912, Dec. 2019, doi: 10.18517/ijaseit.9.6.10226.

[9]. G.Ravikumar, G.Ramachandra, and K.Nagamani,”An Efficient Feature Selection System Integrating SVM with Genetic Algorithm for Large Medical Datasets.*IJARCSSE*, vol. 4.no. 2,pp. 272-277, Mar. 2014.

[10]. A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Feature Subset Selection Problem using Wrapper Approach in Supervised Learning,” *IJCA*, vol. 1, no. 7, pp. 13–17, Feb. 2010, doi: 10.5120/169-295.

[11]. N. M. Hewahi and E. A. Alashqar, “Wrapper Feature Selection based on Genetic Algorithm for Recognizing Objects from Satellite Imagery:,” *Journal of Information Technology Research*, vol. 8, no. 3, pp. 1–20, Jul. 2015, doi: 10.4018/JITR.2015070101.

[12]. J. Huang, Y. Cai, and X. Xu, “A hybrid genetic algorithm for feature selection wrapper based on mutual information,” *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1825–1844, Oct. 2007, doi: 10.1016/j.patrec.2007.05.011.

[13]. Y. Choi, A. Darwiche, and G. Van Den Broeck, “Optimal Feature Selection for Decision Robustness in Bayesian Networks,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, Aug. 2017, pp. 1554–1560. doi: 10.24963/ijcai.2017/215.

[14]. M. R., A. Banu.W, and D. Mavaluru, “An efficient feature selection algorithm for health care data analysis,” *Bulletin EEI*, vol. 9, no. 3, pp. 877–885, Jun. 2020, doi: 10.11591/eei.v9i3.1744.

[15]. S. Oreski and G. Oreski, “Genetic algorithm-based heuristic for feature selection in credit risk assessment,” *Expert Systems with Applications*, vol. 41, no. 4, pp. 2052–2064, Mar. 2014, doi: 10.1016/j.eswa.2013.09.004.

[16]. S. Mirjalili, Ed., *Handbook of moth-flame optimization algorithm: variants, hybrids, improvements, and applications*, First edition. in Advances in metaheuristics. Boca Raton: CRC Press, Taylor & Francis Group, 2023.

[17]. S. Mir and Sunanda, “‘Heart Disease Prediction and Severity Level Classification’: A Machine Learning approach with Feature Selection technique,” in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India: IEEE, Jul. 2023, pp. 1–7. doi: 10.1109/ICCCNT56998.2023.10308175.